Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 439–441, January 1993
Biochemistry

Protein folding—what's the question?

(protein stability/conformational specificity)

EATON E. LATTMAN* AND GEORGE D. ROSE†

*Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, 725 N. Wolfe Street, Baltimore, MD 21205; and †Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, Box 8231, 660 S. Euclid Avenue, St. Louis, MO 63110

Communicated by Christian B. Anfinsen, October 8, 1992

ABSTRACT **The folding reactions of many small, globular proteins exhibit two-state kinetics, in which the folded and unfolded states interconvert readily without observable intermediates. Typically, the free energy difference, ΔG, between the native and denatured states of such a protein is quite small, lying in the range of approximately −5 to −15 kcal/mol. We point out that, under these circumstances, a population of native-like molecules will persist, even in the presence of mutations sufficiently destabilizing to change the sign of ΔG. Therefore, it is not energy *per se* that determines conformation. A corollary to this argument is that specificity—not stability—would be the more informative focus in future folding studies.**

A protein molecule adopts its unique, three-dimensional equilibrium structure spontaneously under physiological conditions in many (1), if not all (2), cases. This native structure can be denatured readily by elevated temperature or perturbing solvents, both of which induce chain disorder but leave covalent bonds intact (3, 4). The transition of the polypeptide chain from a disordered, nonnative state to the ordered, native state is called *protein folding*. It is well established (5) that the information necessary to drive this reversible disorder ↔ order transition is encrypted solely within the linear amino acid sequence; hence, the structure of a protein is implicit in the gene that encodes it.

Despite intense research, a generalized mechanistic understanding of the folding transition remains obscure, and the *protein folding problem*—to understand how the amino acid sequence specifies the three-dimensional structure—has yet to be solved. Driven by the promise of protein engineering (6) and the prospect of the human genome project (7), this problem has emerged as one of the 20th century's most significant scientific challenges. Recently, the so-called inverse folding problem—prediction of sequences that can adopt a given fold—has also received much attention (8–11). In the following, we examine some current thinking about the folding problem and propose an additional paradigm.

The Thermodynamic Point of View

The folding problem was first recognized more than half a century ago (12). Since Kauzmann (13), this has been the central question: what are the determinants of protein stability (14)? This question is epitomized in the thermodynamic hypothesis of Anfinsen, which asserts that the folded state of a globular protein resides at a global minimum of free energy (1).

Thermodynamic measurements of proteins have been particularly informative, due in large part to the observation that the folding transition is two-state, or nearly two-state, with a negligible population of stable intermediates in most cases (15). That is, at the midpoint of the folding transition, half the molecules are folded (N) and half are unfolded (U). This fact of nature simplifies protein thermodynamics by allowing the folding reaction to be represented as U ↔ N, with equilibrium constant K_{eq} = [N]/[U]. The difference in free energy between the native and unfolded states is then given by $\Delta G = -RT \ln K_{eq}$, where R is the gas constant and T is the absolute temperature. For globular proteins, typical values of ΔG are quite small; they lie in the range of −5 to −15 kcal/mol (16).

The Structural Point of View

We now highlight a more structurally related question: what are the determinants of protein conformation? This ostensibly modest change in emphasis results in a dramatic shift in focus and interpretation.

There is a profound organizing aspect of two-state behavior that is often overlooked: *proteins are either folded in a native-like manner or not folded at all.* As a familiar example, lysozyme and ribonuclease have similar molecular weights and compositions but very different folds; mutations that destabilize lysozyme push the folding equilibrium toward unfolded lysozyme, not toward folded ribonuclease. At equilibrium, the reaction U ↔ N represents a partitioning between an ensemble that lacks some property indicative of the folded state and a unique native structure that exhibits that property.

Mutational studies afford a versatile means to probe protein structure (17). Reports from many recent studies (18–24) are well summarized as follows: single-site mutations generally result in small but measurable changes in structure, function, or stability, but they do not affect two-state behavior materially. In the usual thermodynamic interpretation of such experiments, if a single mutation destabilizes a protein by 1–2 kcal/mol (a typical change) and if the difference in free energy between native and denatured states is ≈8 kcal/mol (a typical ΔG), then a few such mutations might be expected to destabilize the molecule altogether (i.e., $\Delta G > 0$).

Logically, the persistence of two-state behavior implies that a population of native-like molecules will exist, even in the presence of the most destabilizing mutations. Thus, the overall conformation of a protein is independent of its stability. This conclusion is no mere exercise in logic. For example, excision of a protein fragment is expected to be radically destabilizing. Yet, two decades ago, work of Sachs *et al.* (25) showed that antibodies raised against intact staphylococcal nuclease could still recognize a short, isolated fragment of the polypeptide.

To emphasize this point, suppose that several destabilizing mutations were introduced, shifting the equilibrium constant so that $\Delta G = +5.5$ kcal/mol. Assuming two-state behavior at physiological RT, only 1 molecule in ≈10,000 would adopt the native conformation; the remaining 9999 would be unfolded. Despite greatly diminished stability, the mechanistic question about determinants of conformation persists; that is, why does the rare folded species still adopt a native-like fold?

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

In essence, there are two separable questions, one thermodynamic, the other structural: (*i*) why do proteins fold at all, and (*ii*) what is the source of *conformational specificity* (i.e., the stereochemical code that directs a protein to adopt its native fold in preference to some other)? That is, why does lysozyme adopt the lysozyme fold and not the ribonuclease fold? Our current understanding of factors that contribute to stability, as analyzed in work from Kauzmann (13) to Dill (26), may have already answered the first question. We turn now to the second.

What is the origin of conformational specificity? One attractive candidate has been internal packing. Globular proteins are known to have mean packing densities reminiscent of solids, a consequence of the exquisite complementarity between interior side chains, which fit together like pieces of a three-dimensional jigsaw puzzle (27). This experimental fact can be interpreted to mean that protein conformation is linked tightly to internal packing. According to this interpretation, lysozyme adopts the lysozyme fold and not the ribonuclease fold because efficient internal packing can be achieved only in the former case.

Earlier, we tested this view by analyzing patterns of complementarity within proteins of known structure (28). Our analysis led to the surprising conclusion that efficient packing is readily attained among clusters of the naturally occurring hydrophobic amino acid residues. The result implies that good packing—an undeniable experimental fact—is not the major determinant of conformation. Of course, given that the protein is folded, the molecular interior is expected to be well-packed. Were it not, then dispersion forces would favor denaturation, since presumably efficient packing could be achieved between protein and solvent in the denatured state. However, the fact that complementary internal packing exists need not imply that packing *per se* determines conformation; and, indeed, our analysis, together with many mutational experiments (18–24), indicates that, in general, packing is not the dominant factor.

This is not to say that alternative packing arrangements are isoenergetic. Indeed, mutations usually lead to measurable changes—most often a decrease—in stability (18–24).

In short, while packing *does* modulate the conformational equilibrium between folding and unfolding (question *i*), it *does not* appear to be the primary source of conformational specificity (question *ii*). What then is?

A Stereochemical Code for Protein Folding

The determinants that control overall conformation could be encoded within the amino acid sequence in either of two extreme formats: centralized or distributed. In centralized control, where several discrete sites ("tender spots") specify the fold, structural integrity could be expected to withstand most mutations, but alteration of a tender spot would result in conformational "catastrophe." In distributed control, conformation would be determined—and most probably overdetermined—by information throughout the sequence, and in this case information needed to specify the fold would be expected to survive alteration or deletion of a few residues. The mutational experiments discussed previously (18–24), in which typical residue substitutions range over the entire molecule and have little impact on the overall fold, are more suggestive of distributed control.

It is plausible that conformational specificity is imposed through a redundant stereochemical code that arises from the interplay between the shape and polarity of residue side chains and secondary structure conformation. Initial evidence for such a code is now emerging in the case of helices. Capping of helix termini involves residue polarity (29–31), whereas residue preferences at central positions of the helix are governed largely by side chain shape, in the form of conformational restrictions imposed by the bulky helix backbone (32). For example, valine can populate all three side chain conformers (gauche$^+$, gauche$^-$, trans) in the unfolded state, but this β-branched side chain is restricted to essentially one conformer (trans) in a helix (32). The corresponding energy loss ($T\Delta S$), due solely to the reduction in side chain configurational entropy, is $\approx RT$ ln 3, approximating physiological RT. Such energy terms cause residues to favor one type of secondary structure over others, and, while individually modest, they are substantial in the aggregate.

This postulated stereochemical code, with the protein fold determined by a large number of individually small interactions, represents a robust, distributed folding mechanism. Further, in such a code, mediation of conformational specificity would not be coupled tightly to folding stability, although factors that stabilize a given element of secondary structure would, in general, be expected to stabilize the entire molecule (R. J. Pinker, G.D.R., and N. R. Kallenbach, unpublished results).

In sum, we propose that the derivation of a reliable strategy to predict structure from sequence will depend critically upon elucidation of the stereochemical code that underlies conformational specificity.

We are grateful to Gary Ackers, Mario Amzel, Jeremy Berg, Carl Frieden, Timothy Lohman, Susan Panny, Judith Robertson, Richard Wolfenden, and Bruno Zimm for their critical reading of this manuscript. This work was supported by National Institutes of Health Grants GM 35171 (E.E.L.) and GM29458 (G.D.R.).

1. Anfinsen, C. B. (1973) *Science* **181,** 223–230.
2. Gething, M.-J. & Sambrook, J. (1992) *Nature (London)* **355,** 33–45.
3. Tanford, C. (1968) *Adv. Prot. Chem.* **23,** 121–282.
4. Kim, P. S. & Baldwin, R. L. (1982) *Annu. Rev. Biochem.* **51,** 459–489.
5. Anfinsen, C. B. & Scheraga, H. A. (1975) *Adv. Prot. Chem.* **29,** 205–300.
6. Oxender, D. L. & Fox, C. F., eds. (1987) *Protein Engineering* (Liss, New York).
7. Congress of the United States, Office of Technology Assessment (1988) *Mapping Our Genes* (U.S. Government Printing Office, Washington, DC).
8. Drexler, K. E. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 5275–5278.
9. Pabo, C. (1983) *Nature (London)* **301,** 200.
10. Ponder, J. W. & Richards, F. M. (1987) *J. Mol. Biol.* **193,** 775–791.
11. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) *Science* **253,** 164–170.
12. Mirsky, A. E. & Pauling, L. (1936) *Proc. Natl. Acad. Sci. USA* **22,** 439–447.
13. Kauzmann, W. (1959) *Adv. Prot. Chem.* **14,** 1–64.
14. Privalov, P. L. (1979) *Adv. Prot. Chem.* **33,** 167–241.
15. Schellman, J. A. (1987) *Annu. Rev. Biophys. Biophys. Chem.* **16,** 115–137.
16. Pace, C. N. (1990) *Trends Biochem. Sci.* **15,** 14–17.
17. Alber, T. (1989) *Annu. Rev. Biochem.* **58,** 765–798.
18. Estell, D. A., Graycar, T. P., Miller, J. V., Powers, D. B., Burnier, D. B., Ng, P. G. & Wells, J. A. (1986) *Science* **233,** 659–663.
19. Matthews, B. W. (1987) *Biochemistry* **26,** 6885–6888.
20. Matouschek, A., Kellis, J. T., Serrano, L. & Fersht, A. R. (1989) *Nature (London)* **340,** 122–126.
21. Sondek, J. E. & Shortle, D. (1990) *Proteins Struct. Funct. Genet.* **7,** 299–305.
22. Sandberg, W. S. & Terwilliger, T. C. (1990) *Science* **245,** 54–57.
23. Lim, W. A. & Sauer, R. T. (1991) *J. Mol. Biol.* **219,** 359–376.
24. Eriksson, A. E., Basse, W. A., Zhang, X.-J., Heinz, D. W., Blaber, M., Baldwin, E. P. & Matthews, B. W. (1992) *Science* **255,** 178–183.
25. Sachs, D. H., Schechter, A. N., Eastlake, A. & Anfinsen, C. B. (1972) *J. Immunol.* **109,** 1300–1310.
26. Dill, K. A. (1990) *Biochemistry* **29,** 7133–7155.

27. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6,** 151–176.
28. Behe, M. J., Lattman, E. E. & Rose, G. D. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 4195–4199.
29. Shoemaker, K. R., Kim, P. S., York, E. J., Stewart, J. M. & Baldwin, R. L. (1987) *Nature (London)* **326,** 563–567.
30. Presta, L. G. & Rose, G. D. (1988) *Science* **240,** 1632–1641.
31. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648–1652.
32. Creamer, T. P. & Rose, G. D. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 5937–5941.